Davide Belli
11887532
davidebelli95@gmail.com

Homework Assignment 1
Machine Learning 2, 17/18

2018-04-10

In this Homework, I discussed possible interpretations with Gabriele Cesa and Pascal Esser, Andrew Skliar, Gabriele Bani.

1 Problem 1

Assumption: **x** and **z** are vectors drawn from independent distributions

$$\begin{aligned}\mathbf{E}[\mathbf{y}] &= \mathbf{E}[\mathbf{x}+\mathbf{z}] = \\ &= \mathbf{E}[\mathbf{x}] + \mathbf{E}[\mathbf{z}] = \\ &= \mu_{\mathbf{x}} + \mu_{\mathbf{z}}\end{aligned}$$

$$\begin{aligned}Cov(\mathbf{y},\mathbf{y}) &= Cov(\mathbf{x}+\mathbf{z},\mathbf{x}+\mathbf{z}) \\ &= \mathbf{E}[(\mathbf{x}+\mathbf{z})(\mathbf{x}+\mathbf{z})^T] - \mathbf{E}[(\mathbf{x}+\mathbf{z})]\mathbf{E}[(\mathbf{x}+\mathbf{z})\end{aligned}$$

Let's consider now for simplicity the element in the matrix in row i, column j:
$Cov(\mathbf{y},\mathbf{y})_{i,j}$

$$\begin{aligned}Cov(\mathbf{y},\mathbf{y})_{i,j} &= \mathbf{E}[(\mathbf{x}_i+\mathbf{z}_i)(\mathbf{x}_j+\mathbf{z}_j)] - \mathbf{E}[(\mathbf{x}_i+\mathbf{z}_i)]\mathbf{E}[(\mathbf{x}_j+\mathbf{z}_j) = \\ &= \mathbf{E}[\mathbf{x}_i\mathbf{x}_j] + \mathbf{E}[\mathbf{z}_i\mathbf{z}_j] + \mathbf{E}[\mathbf{z}_i\mathbf{z}_j] + \mathbf{E}[\mathbf{x}_j\mathbf{z}_i] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{x}_j] - \mathbf{E}[\mathbf{z}_i]\mathbf{E}[\mathbf{z}_j] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{z}_j] - \mathbf{E}[\mathbf{x}_j]\mathbf{E}[\mathbf{z}_i] = \\ &= [\mathbf{E}[\mathbf{x}_i\mathbf{x}_j] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{x}_j]] + [\mathbf{E}[\mathbf{z}_i\mathbf{z}_j] - \mathbf{E}[\mathbf{z}_i]\mathbf{E}[\mathbf{z}_j]] + [\mathbf{E}[\mathbf{z}_i\mathbf{z}_j] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{z}_j]] + [\mathbf{E}[\mathbf{x}_j\mathbf{z}_i] - \mathbf{E}[\mathbf{x}_j]\mathbf{E}[\mathbf{z}_i]] = \\ &= Cov(\mathbf{x},\mathbf{x})_{i,j} + Cov(\mathbf{z},\mathbf{z})_{i,j} + Cov(\mathbf{x},\mathbf{z})_{i,j} + Cov(\mathbf{z},\mathbf{x})_{i,j} = \\ &= (\Sigma_{\mathbf{x}})_{i,j} + (\Sigma_{\mathbf{z}})_{i,j} + 0 + 0\end{aligned}$$

Then:

$$Cov(\mathbf{y},\mathbf{y}) = \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{z}}$$

2 Problem 2

2.1 Question 1

$$\begin{aligned}p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \\ &= \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \\ &= \prod_{i=1}^{N} (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\end{aligned}$$

$$\begin{aligned}LL(p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mu},\boldsymbol{\Sigma})) &= \log\left(\prod_{i=1}^{N} (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\right) = \\ &= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\end{aligned}$$

2.2 Question 2

$$
\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) &= p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) \\
&= \frac{\prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})}{p(\boldsymbol{\mathcal{X}})} = \\
&= \frac{\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})}{\int_{\boldsymbol{\mu'}} \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu'}, \boldsymbol{\Sigma})\mathcal{N}(\boldsymbol{\mu'}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})d\boldsymbol{\mu'}}
\end{aligned}
$$

2.3 Question 3

$$
\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) &= \frac{\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})}{p(\boldsymbol{\mathcal{X}})} \\
&consider \\
\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma_l})
\end{aligned}
$$

where, for product of i.i.d gaussians,

$$
\boldsymbol{\Sigma_l} = \frac{1}{N}\boldsymbol{\Sigma}
$$

then, considering only the numerator of the posterior:

$$
\begin{aligned}
\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma_l})\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) &= \\
(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma_l}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i}^{N}(\boldsymbol{x_i}-\boldsymbol{\mu})^T\boldsymbol{\Sigma_l}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu})\right) & \\
(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma_0}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\right) &= \\
K \exp\left(-\frac{1}{2}\sum_{i}^{N}(\boldsymbol{x_i}-\boldsymbol{\mu})^T\boldsymbol{\Sigma_l}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\right) &
\end{aligned}
$$

Where K is the normalizing constant independent from μ, therefore, we can look at the exponent of the resulting at the exponential function to put it in form of the exponent of a new Gaussian function.

$$
\begin{aligned}
-\frac{1}{2}\sum_{i}^{N}(\boldsymbol{x_i}-\boldsymbol{\mu})^T\boldsymbol{\Sigma_l}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0}) &= \\
\frac{1}{2}(\boldsymbol{\mu}^T(\boldsymbol{\Sigma_l}^{-1}+\boldsymbol{\Sigma_0}^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^T(\boldsymbol{\Sigma_l}^{-1}\sum_{i}^{N}\boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}) + \boldsymbol{\mu_0}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \sum_{i}^{N}\boldsymbol{x_i}^T\boldsymbol{\Sigma_l}^{-1}\boldsymbol{x_i}) &
\end{aligned}
$$

then, we can see that the new covariance matrix in the first term must have the form:

$$
\boldsymbol{\Sigma_N}^{-1} = \boldsymbol{\Sigma_l}^{-1} + \boldsymbol{\Sigma_0}^{-1}
$$

and, for derivation of the products in the multivariate gaussian, considering the covariance matrix that we just found and the second term in the previous result:

$$
2\boldsymbol{\mu}^T\boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N} = 2\boldsymbol{\mu}^T(\boldsymbol{\Sigma_l}^{-1}\sum_{i}^{N}\boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0})
$$

$$\boldsymbol{\Sigma_N}\boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N} = \boldsymbol{\Sigma_N}(\boldsymbol{\Sigma_l}^{-1}\sum_i^N \boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0})$$
$$\boldsymbol{\mu_N} = \boldsymbol{\Sigma_N}(\boldsymbol{\Sigma_l}^{-1}\sum_i^N \boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0})$$

Finally, we can put back together these parts in the original form, using:
$K' = K\exp(+\boldsymbol{\mu_0}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \sum_i^N \boldsymbol{x_i}^T\boldsymbol{\Sigma_l}^{-1}\boldsymbol{x_i})$ Since that part of the argument of the exponent is independent from μ.

$$\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) &= \frac{\prod_{i=1}^N \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})}{\int_{\boldsymbol{\mu'}}\prod_{i=1}^N \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu'}, \boldsymbol{\Sigma})\mathcal{N}(\boldsymbol{\mu'}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})d\boldsymbol{\mu'}} \\
&= \frac{K\exp\left(-\frac{1}{2}\sum_i^N(\boldsymbol{x_i}-\boldsymbol{\mu})^T\boldsymbol{\Sigma_l}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\right)}{\int_{\boldsymbol{\mu'}} K\exp\left(-\frac{1}{2}\sum_i^N(\boldsymbol{x_i}-\boldsymbol{\mu'})^T\boldsymbol{\Sigma_l}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu'}) - \frac{1}{2}(\boldsymbol{\mu'}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu'}-\boldsymbol{\mu_0})\right)d\boldsymbol{\mu'}} \\
&= \frac{K'\exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}^T(\boldsymbol{\Sigma_l}^{-1}+\boldsymbol{\Sigma_0}^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^T(\boldsymbol{\Sigma_l}^{-1}\sum_i^N\boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0})\right)\right)}{\int_{\boldsymbol{\mu'}} K'\exp\left(-\frac{1}{2}\left(\boldsymbol{\mu'}^T(\boldsymbol{\Sigma_l}^{-1}+\boldsymbol{\Sigma_0}^{-1})\boldsymbol{\mu'} - 2\boldsymbol{\mu'}^T(\boldsymbol{\Sigma_l}^{-1}\sum_i^N\boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0})\right)\right)d\boldsymbol{\mu'}} \\
&= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_N}, \boldsymbol{\Sigma_N}) \\
&= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\Sigma_N}(\boldsymbol{\Sigma_l}^{-1}\sum_i^N\boldsymbol{x_i} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}), \boldsymbol{\Sigma_l}^{-1}+\boldsymbol{\Sigma_0}^{-1})
\end{aligned}$$

2.4 Question 4

$$\boldsymbol{\mu}_{MAP}:$$
$$\frac{\partial}{\partial\boldsymbol{\mu}}\log p(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) = 0$$
$$\frac{\partial}{\partial\boldsymbol{\mu}}\Big(-\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^N(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$
$$-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\Big) = 0$$
$$\frac{\partial}{\partial\boldsymbol{\mu}}\Big(-\frac{1}{2}\sum_{i=1}^N(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0}) + K\Big) = 0$$
$$-\boldsymbol{\Sigma}^{-1}\sum_{i=1}^N(\boldsymbol{x}_i-\boldsymbol{\mu}) - \boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0}) = 0$$
$$-\boldsymbol{\Sigma}^{-1}\sum_{i=1}^N(\boldsymbol{x}_i) + N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu} = 0$$
$$N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}^{-1}\sum_{i=1}^N(\boldsymbol{x}_i) + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}$$
$$\boldsymbol{\mu}_{MAP} = \frac{\boldsymbol{\Sigma}^{-1}\sum_{i=1}^N(\boldsymbol{x}_i) + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}}{N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma_0}^{-1}}$$

3 Problem 3

3.1 Question 1

With n_0 =number of tail and n_1 = number of heads:

$$\boldsymbol{\mu}_{MLE}:$$

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \log p(\mathcal{X}|\mu) &= 0 \\
\frac{\partial}{\partial \mu} \log p(\mathcal{X}|\mu) &= \frac{\partial}{\partial \mu} \log \prod_{i=1}^{N} \mu_i^x (1-\mu)^{(1-x_i)} \\
&= \frac{\partial}{\partial \mu} \sum_{i=1}^{N} x_i \log \mu + (1-x_i) \log(1-\mu) \\
&= \frac{x_i}{\mu} - \frac{x_i}{1-\mu} \\
(1-\mu) n_0 &= \mu n_1 \\
\mu_{MLE} &= \frac{n_1}{n_0+n_1} = \frac{3}{0+3} = 1
\end{aligned}
$$

3.2 Question 2

$$
\begin{aligned}
&\mu_{MAP}: \\
&\frac{\partial}{\partial \mu} \log p(\mathcal{X}|\mu) p(\mu|a,b) = 0 \\
\frac{\partial}{\partial \mu} \log p(\mu|a,b) &= \frac{\partial}{\partial \mu} \log \prod_{i=1}^{N} \mu_i^x (1-\mu)^{(1-x_i)} \frac{\mu^{a-1}(1-\mu)^{b-1}}{B(a,b)} \\
&= \frac{\partial}{\partial \mu} \sum_{i=1}^{N} \big(x_i \log \mu + (1-x_i)\log(1-\mu)\big) + (a-1)\log \mu + (b-1)\log(1-\mu) - logB(a,b) \\
&= \frac{x_i}{\mu} - \frac{x_i}{1-\mu} + \frac{a-1}{\mu} - \frac{b-1}{1-\mu} \\
\mu_{MAP} &= \frac{n_1+a-1}{n_0+n_1+a+b-2} \\
&= \frac{a+2}{a+b+1}
\end{aligned}
$$

While the posterior distribution will be: $\mu \sim Beta(a+n_1, b+n_0) = Beta(a+3, b)$

3.3 Question 3

We can show how the MAP can be represented as a linear combination of the prior mean (weighted $\frac{H}{H+K}$) and the μ_{MLE} (weighted $\frac{K}{H+K}$), and thus it always lies inbetween them. Consider the prior mean:

$$
\mathbf{E}[p(\mu|a,b)] = \frac{a}{a+b}
$$

Then, consider $K = n_0 + n_1$ and $H = a + b$:

$$
\begin{aligned}
\frac{H}{H+K}\mathbf{E}[p(\mu|a,b)] + \frac{K}{H+K}\mu_{MLE} &= \\
\frac{H}{H+K}\frac{a}{a+b} + \frac{K}{H+K}\frac{n_1}{n_0+n_1} &= \\
\frac{a+b}{n_0+n_1+a+b}\frac{a}{a+b} + \frac{n_0+n_1}{n_0+n_1+a+b}\frac{n_1}{n_0+n_1} &= \\
\frac{a}{n_0+n_1+a+b} + \frac{n_1}{n_0+n_1+a+b} &= \\
\frac{n_0+a}{n_0+n_1+a+b} &= \\
\mathbf{E}[p(\mu|\mathcal{X},a,b)]
\end{aligned}
$$

4 Problem 4

Exponential family:

$$p(x|\boldsymbol{\eta}) = b(x)\exp(\boldsymbol{\eta}^T T(x) - A(\boldsymbol{\eta})]$$

4.1 Question 1

4.1.1 Poisson

$$\begin{aligned} Poisson(k|\lambda) &= \frac{\lambda^k e^{-\lambda}}{k!} \\ &= \frac{1}{k!} e^{k\log\lambda - \lambda} \end{aligned}$$

$$\begin{aligned} b(x) &= \frac{1}{k!} \\ \eta &= \log\lambda \\ \lambda &= e^\eta \\ T(k) &= k \\ A(\eta) &= \lambda = e^\eta \end{aligned}$$

4.1.2 Gamma

$$\begin{aligned} Gamma(\tau|a,b) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} \\ &= e^{-\log\Gamma(a) + a\log b + (a-1)\log\tau - b\tau} \end{aligned}$$

$$\begin{aligned} b(x) &= 1 \\ \eta_1 &= a - 1 \\ \eta_2 &= -b \\ a &= \eta + 1 \\ b &= -\eta_2 \\ T_1(\tau) &= \log\tau \\ T_2(\tau) &= \tau \\ A(\eta_1,\eta_2) &= \log\Gamma(a) - a\log b \\ &= \log\Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2) \end{aligned}$$

4.1.3 Cauchy

Cauchy distribution cannot be expressed in exponential form, and indeed it is not part of the exponential family. To prove this, is enough to consider that members of the exponential family must have finite moments, while Cauchy distribution do not have finite moments for any order.

4.1.4 Von Mises

$$\begin{aligned} VonMises(x|k,\mu) &= \frac{1}{2\pi I_0(k)} e^{k\cos(x-\mu)} \\ &= \frac{1}{2\pi I_0(k)} e^{k\cos x\cos\mu + ksinx\sin\mu} \\ &= \exp\left(k\cos x\cos\mu + ksinx\sin\mu - \log(2\pi I_0(k))\right) \end{aligned}$$

$$\begin{aligned} b(x) &= 1 \\ \eta_1 &= k\cos\mu \\ \eta_2 &= k\sin\mu \end{aligned}$$

$$
\begin{aligned}
\frac{\eta_2}{\eta_1} &= \frac{\sin\mu}{\cos\mu} \\
\mu &= \arctan\frac{\eta_2}{\eta_1} \\
\sqrt{\cos^2\mu + \sin^2\mu} &= \sqrt{(\frac{\eta_1}{k})^2 + (\frac{\eta_2}{k})^2} = 1 \\
k &= \sqrt{\eta_1^2 + \eta_2^2} \\
T(x) &= [\cos x, \sin x]^T \\
A(\eta_1, \eta_2) &= -\log(2\pi I_0(k)) \\
&= -\log(2\pi I_0(\sqrt{\eta_1^2 + \eta_2^2}))
\end{aligned}
$$

4.2 Question 2

4.2.1 Poisson

Using Bishop (2.226):

$$
\begin{aligned}
\mathbf{E}[T(k)] &= \mathbf{E}[k] \\
&= \nabla_\eta A(\eta) \\
&= \nabla_\eta e^\eta \\
&= e^\eta \\
&= \lambda
\end{aligned}
$$

$$
\begin{aligned}
Var[T(k)] &= \mathbf{E}^2[k] \\
&= \nabla_\eta^2 A(\eta) \\
&= \nabla_\eta^2 e^\eta \\
&= e^\eta \\
&= \lambda
\end{aligned}
$$

Deriving the second moment (as request in the task) withouth the generator of moments:

$$
\begin{aligned}
\mathbf{E}^2[p(k|\lambda)] &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} \\
&= \sum_{k=1}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} \\
&= \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} \\
&= \lambda \sum_{k=0}^{\infty} (k+1) \frac{\lambda^k e^{-\lambda}}{k!} \\
&= \lambda \mathbf{E}[p(k+1|\lambda)] \\
&= \lambda \mathbf{E}[p(k|\lambda) + 1] \\
&= \lambda(\lambda + 1) \\
&= \lambda^2 + \lambda
\end{aligned}
$$

4.2.2 Gamma

Using Bishop (2.226):

$$
\begin{aligned}
\mathbf{E}[T_2(\tau)] &= \mathbf{E}[\tau] \\
&= \nabla_{\eta_2} A(\eta_1, \eta_2) \\
&= \nabla_{\eta_2} \log\Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2) \\
&= \frac{\eta_1 + 1}{-\eta_2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{a-1+1}{b} \\
&= \frac{a}{b}
\end{aligned}
$$

$$
\begin{aligned}
Var[T_2(\tau)] &= \mathbf{E}^2[\tau] \\
&= \nabla^2_{\eta_2} A(\eta_1, \eta_2) \\
&= \nabla^2_{\eta_2} \log\Gamma(\eta_1+1) - (\eta_1+1)\log(-\eta_2) \\
&= \nabla_{\eta_2} \frac{\eta_1+1}{-\eta_2} \\
&= \frac{\eta_1+1}{\eta_2^2} \\
&= \frac{a-1+1}{b^2} \\
&= \frac{a}{b^2}
\end{aligned}
$$

If we want to actually derive the second moment (not the Variance), we find:

$$
\begin{aligned}
\mathbf{E}^2[p(\tau|a,b)] &= \int \frac{1}{\Gamma(a)} b^a \tau^{a+1} e^{-b\tau} d\tau \\
&= \frac{\Gamma(a+2)}{\Gamma(a)b^2} \int \Gamma(\tau|a+2,b) \\
&= \frac{\Gamma(a+2)}{\Gamma(a)b^2} \\
&= \frac{a(a+1)}{b^2}
\end{aligned}
$$

4.3 Question 3

The Poisson distribution has a conjugate prior, indeed, members of the exponential family always have a conjugate prior that can be written in a fixed form. Let's now demonstrate that the Gamma distribution is a conjugate prior for the Poisson.

$$
\begin{aligned}
p(\lambda|k,a,b) &= \frac{p(k|\lambda)p(\tau|a,b)}{p(k)} \\
&= \frac{\frac{\lambda^k e^{-\lambda}}{k!}\frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}}{\int_{\lambda'} \frac{\lambda'^k e^{-\lambda'}}{k!}\frac{1}{\Gamma(a)} b^a \lambda'^{a-1} e^{-b\lambda'} d\lambda'} \\
&= \frac{\lambda^{a+k-1} e^{-(b+1)\lambda}}{\int_{\lambda'} \lambda'^{a+k-1} e^{-(b+1)\lambda'} d\lambda'} \\
&= \frac{\lambda^{a+k-1} e^{-(b+1)\lambda}}{\Gamma(a+k-1)} \\
&= Gamma(\lambda|a+k-1, b+1)
\end{aligned}
$$

5 Problem 5

Using results found in Some Characterizations of the Multivariate t Distribution* , we can use the representation of the multivariate t as:

$$
\begin{aligned}
\boldsymbol{X} &= S^{-1}\boldsymbol{Y} + \boldsymbol{\mu} \\
where & \\
\boldsymbol{X} &\sim T_{\boldsymbol{\nu}}(\boldsymbol{\mu}, \boldsymbol{\sigma}, p) \\
\nu S^2 &\sim \mathcal{X}_\nu{}^2
\end{aligned}
$$

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma})$$

where $\boldsymbol{Y}$ and S^{-1} are independent variables, then

$$\begin{aligned}
\mathbf{E}[\boldsymbol{X}] &= S^{-1}\boldsymbol{Y} + \boldsymbol{\mu} \\
&= \mathbf{E}[S^{-1}\boldsymbol{Y}] + \boldsymbol{\mu} \\
&= \mathbf{E}[S^{-1}]\mathbf{E}[\boldsymbol{Y}] + \boldsymbol{\mu} \\
&= \mathbf{E}[S^{-1}]\mathbf{0} + \boldsymbol{\mu} \\
&= \boldsymbol{\mu}
\end{aligned}$$

Now, we know that by definition, Student-t distribution is unimodal and symmetric. As a consequence, this implies that $Mode[\boldsymbol{X}] = \mathbf{E}[\boldsymbol{X}] = \boldsymbol{\mu}$
Finally, we can use the same representation as above to compute the Variance, also considering the ”Law of total variance”: $Var[X] = \boldsymbol{E}[Var[X|Y]] + Var[\boldsymbol{E}[X|Y]]$ And the conditional probability: $(\boldsymbol{X}|S^2 = s^2) \sim (\boldsymbol{\mu}, s^{-2}\boldsymbol{\Sigma})$
Then we find:

$$\begin{aligned}
Var[\boldsymbol{X}] &= \boldsymbol{E}[Var[X|S^2]] + Var[\boldsymbol{E}[X|S^2]] \\
&= \boldsymbol{E}[s^{-2}\boldsymbol{\Sigma}] + Var[\boldsymbol{\mu}] \\
&= \boldsymbol{E}[\nu(\nu s^2)^{-1}\boldsymbol{\Sigma}] \\
&= \boldsymbol{E}[\nu((\mathcal{X}_\nu^2)^{-1})\boldsymbol{\Sigma}] \\
&= \nu\boldsymbol{E}[((\mathcal{X}_\nu^2)^{-1})]\boldsymbol{\Sigma} \\
&= \frac{\nu\boldsymbol{\Sigma}}{\nu - 2}
\end{aligned}$$

Reminding that the mean of the Inverse Chi-Squared is $\frac{1}{\nu-2}$